Speaker Diarization for Conference Room: The UPC RT09 Evaluation System

Jordi Luque and Javier Hernando

Signal Theory and Communications Department
Technical University of Catalonia (UPC)
Barcelona, Spain

May 29, 2009. Melbourne (Florida), USA

Outline

Introduction

The UPC RT09 Evaluation System

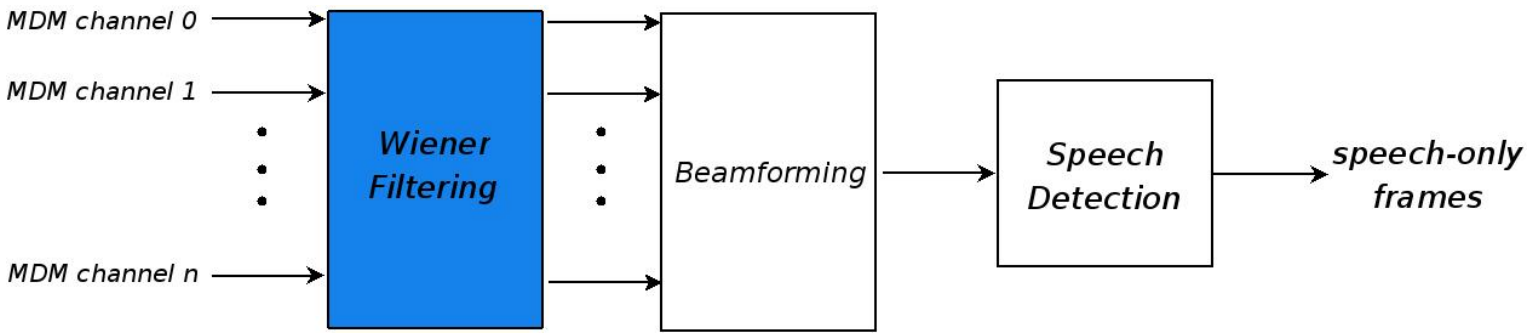
- Perform speaker clustering using a system similar to the system we used in the RT07 evaluation
- Based on ergodic HMM with two streams: MFCC and TDOA
- Top-down merging cluster strategy based on BIC measure among cluster
- Use of a language modelling
- GMM parallel training
- Iterative HMM re-alignment
- Complexity selection

Introduction

Two main blocks

- Signal Enhancement and Speech Detection
  - Noise Reduction applying Wiener filtering
  - Weighted delay and sum beamforming of all available channels
  - Speech frames selection, Speech Segmentation
- Speaker Clustering
  - Ergodic HMM/GMM topology
  - Top-down strategy based on BIC distance among clusters
  - Iterative HMM/GMM re-alignment and automatic complexity selection

Signal Enhancement

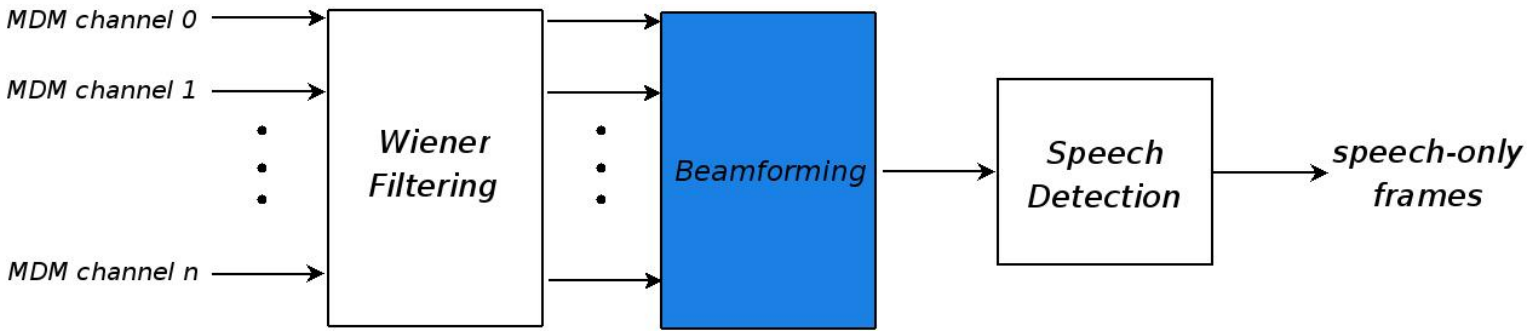


Wiener Filtering

- Wiener filtering for additive noise removal on all of the available MDM channels
- Noise reduction taken from the implementation developed for Aurora2 front-end

A. Adami, L. Burget, S. Dupont, H. Garudadri, F. Grezl, H. Hermansky, P. Jain, S. Kajarekar, N. Morgan, and S. Sivadas, Q̈ualcomm-ICSI-OGI features for ASR; ICSLP 2002

Signal Enhancement

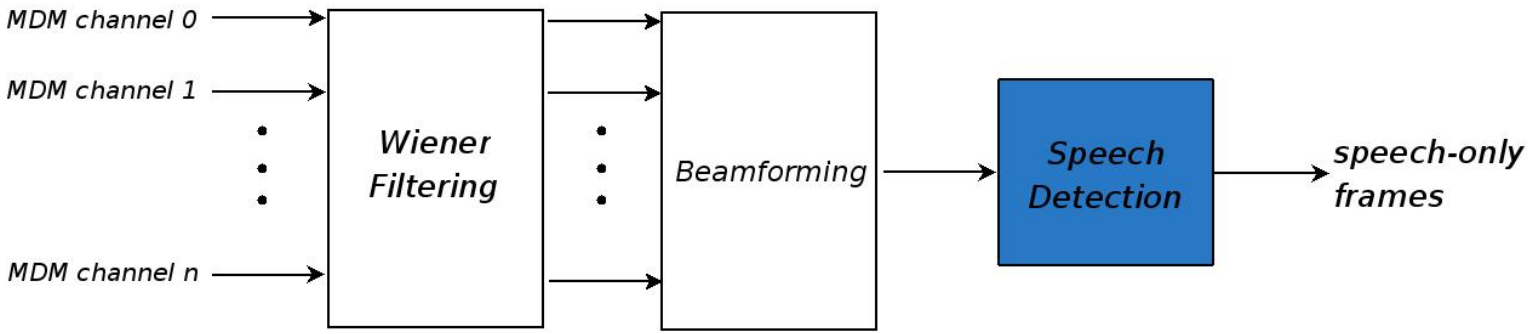


Weighted Delay and Sum Beamforming

- Each noise-reduced channel is processed to obtain a single enhanced channel.
- Weighted delay and sum algorithm BeamformIt (v3.3) with analysis windows of 500ms with shift of 250ms and reference channel selected based on the SNR.

Xavier Anguera, Chuck Wooters and Javier Hernando, "Speaker diarization for multi-party meetings using acoustic fusion", IEEE Automatic Speech Recognition and Understanding Workshop, Puerto Rico, USA, 2005

Speech Detection

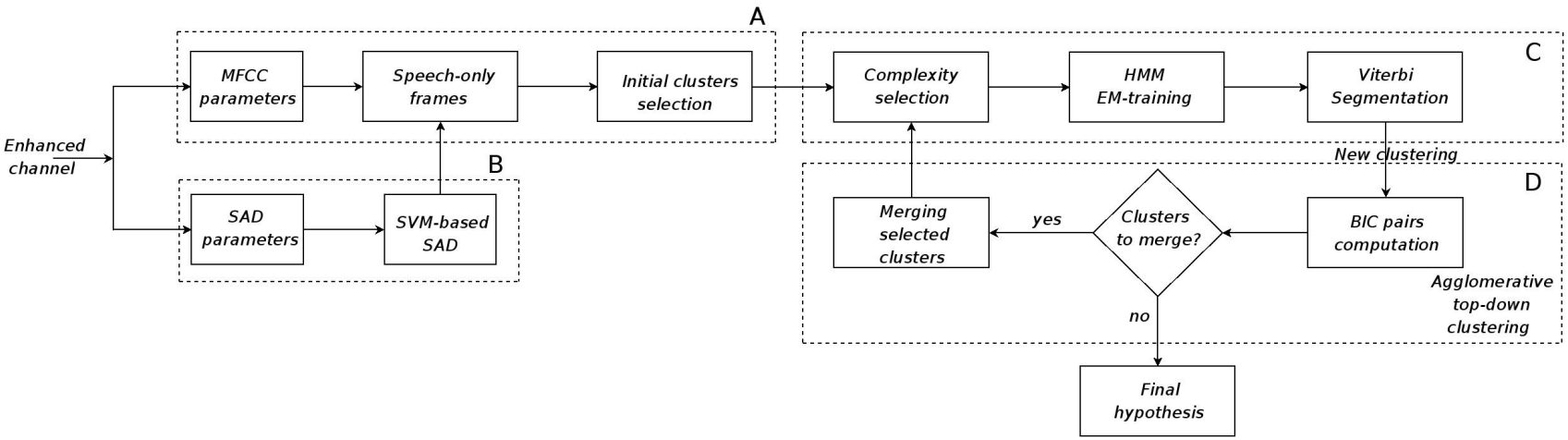


Speech Detection Based on SVM classifier

- The developed system showed a good performance in the last RT SAD and SPKR evaluations
- RT05, RT06, RT07 conference and lecture data was used for training
- Fast training algorithm based on Proximal SVM (PSVM)
- Adjusting Speech-NonSpeech detection in testing stage through the bias b of the separating hyperplane

A. Temko and D. Macho and C. Nadeu, "Enhanced SVM Training for Robust Speech Activity Detection", Proc. ICCASP,2007

Global description



- **A**: MFCC extraction, speech frames selection and clustering initialization
- **B**: Speech detection
- **C**: Complexity selection of the models and HMM/GMM training and clustering realignment
- **D**: Agglomerative clustering based on BIC metric and join stopping criterion with Viterbi score and BIC values

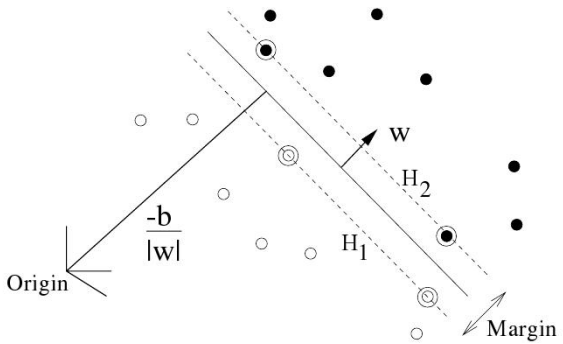
Novelties

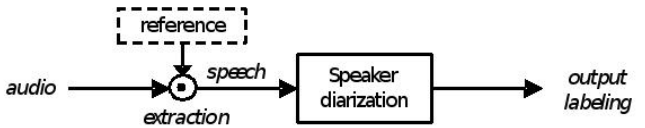
Development

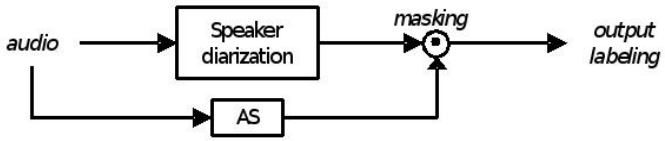
- Rich Transcription evaluation meetings for developing the algorithms.
  - RT05 conference data, RT06 conference and lecture data, RT07 conference data were used to train the SAD.
  - RT06 and RT07 databases were used for assessing and refining the speaker clustering.
- Towards decreasing the runtime of the system while maintaining as much as possible the performance, a look up table has been employed to compute the logarithm function leading to a trade off among speed and accuracy

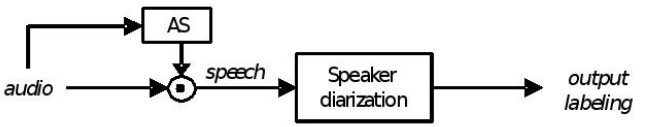
O. Vinyals, G. Friedland, N. Mirgafhori, "Revisiting a basic function on current CPUs: A fast logarithm implementation with adjustable accuracy", ICSI report, June 2007

- **Bias modification in testing**:
  - Trying to adjust the bias of the hyperplane favoring the detection of one class respect the other one
  - It can be noted in the Misses and False Alarm errors
  - Goal: Estimation of the dependence on the SAD performance
- **Three different schemes**:
  - A perfect SAD by means the references (lower performance threshold)
  - Common pre-processing and selection of speech-frames
  - All show data is processed by the system. A masking is applied on the clustering output for discarding the non-speech regions.









SAD Development

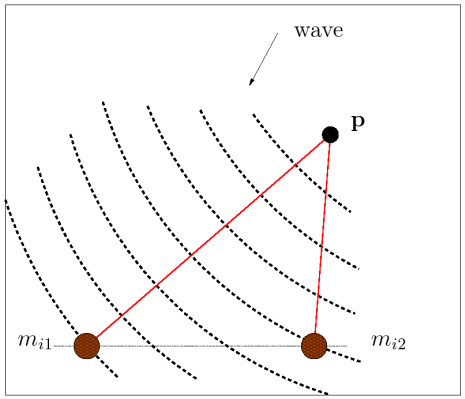
Development results on the RT07 conference data depending on the SVM model's bias b

MISS (%)	FA (%)	SPK ERR (%)	DER (%)
7.0	1.6	7.2	15.79
6.6	1.7	9.5	17.76
6.1	2.0	6.6	14.72
5.7	2.3	4.6	12.61
5.3	2.7	6.7	14.72
5.2	2.8	4.7	12.76
5.0	3.2	5.0	13.1
4.7	3.8	6.3	14.80
4.4	4.7	5.6	14.77
4.2	6.4	5.3	15.94
4.2	6.9	3.4	14.47

System	MISS (%)	FA (%)	SER (%)	DER (%)
Perfect SAD (reference)	3.7	0	6.5	10.27
Post-processing	5.7	2.3	4.5	12.41

TDOA infomation

- **Time Difference of Arrival (TDOA) estimation**:
  - Suitable for speech
  - No limitation in array geometry
  - No assumptions on far field, or narrowband signal

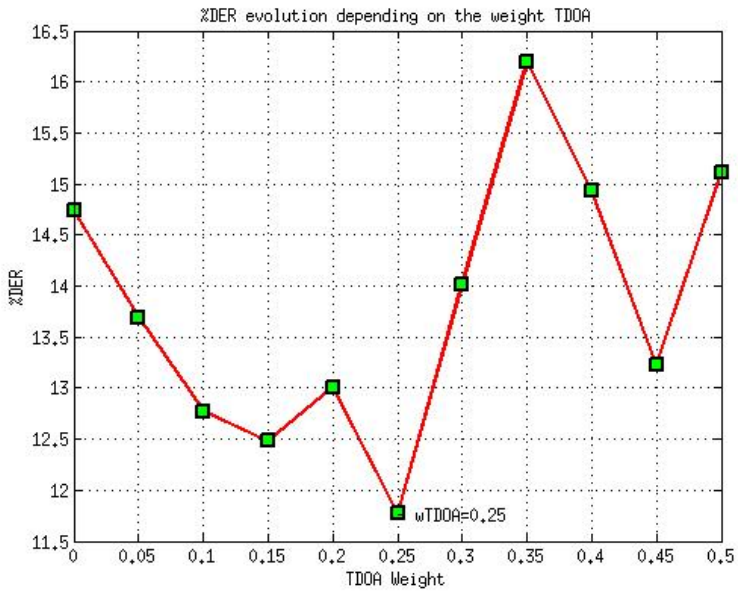


- **Including TDOA information**:
  - Computation of TDOA using a reference channel selected based on SNR features
  - Analysis window of 500ms at a rate of 250ms. Repetition of delays to allow synchronicity with MFCC stream
  - Mixture of Gaussians for fitting the TDOA distribution and weighted linear fusion at score level with the MFCC models.

José M. Pardo, X. Anguera, C. Wooters, "Speaker Diarization for Multiple-Distant-Microphone Meetings Using Several Source of Information", IEEE Transactions on Computers, Vol. 56, No. 9, September 2007

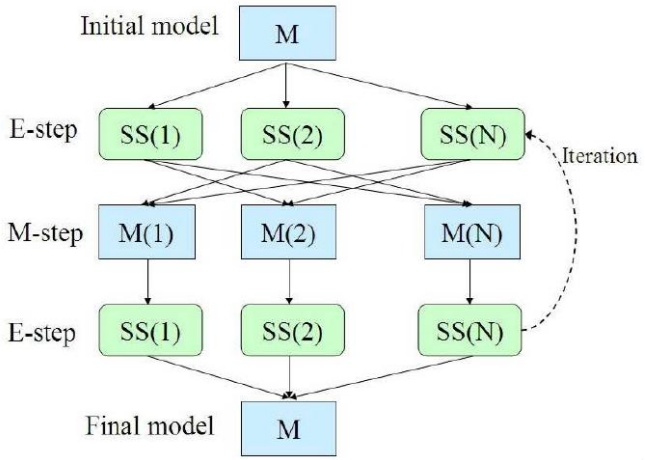
TDOA development

- We found troubles adjusting the weights due the fact TDOAs dimension is dependent on the number of channels. It leads to a higher variance in the dynamic range of the likelihoods among shows. That suggests an specific weight per show and not a global weight. Anyway the use of TDOA obtains a relative improvement of 20%



Cross-Validation EM training

- Avoid overfitting of the data
- Data splitted into N partitions SS_i and conditional probabilities are computed using the initial model M
- Each Model M_i is estimated using $\sum_{j \neq i} SS_j$, SS_i is used as cross-validation data
- Once reached the convergence, the current suficient statistics are used to derive the final model



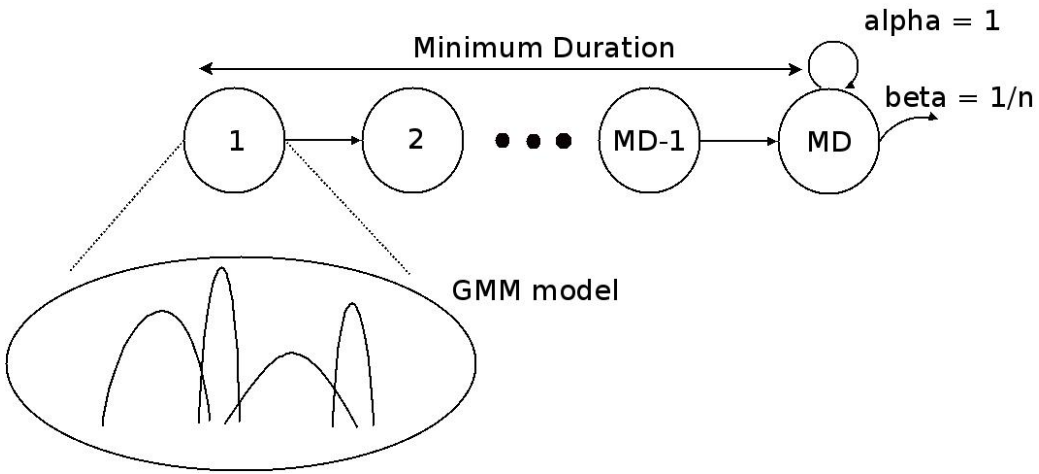
A 5% of total DER relative improvement is obtained

X. Anguera, T. Shinozaki, C. Wooters and J. Hernando, "Model complexity selection and cross-validation EM training for robust speaker diarization", ICASSP'08. Picture courtesy of Xavi Anguera

CV-EM results on RT06-07

Name Show	MISS (%)	FA (%)	SER (%)	DER (%)
CMU_20050912-0900	24.0	1.2	2.4 (3.2)	27.60 (28.32)
CMU_20050914-0900	21.8	1.6	2.1 (2.5)	25.56 (26.01)
CMU_20061115-1030	8.1	3.6	5.8 (4.8)	17.51 (16.45)
CMU_20061115-1530	3.5	3.9	0.7 (0.6)	8.06 (8.00)
EDI_20050216-1051	14.0	7.2	31.9 (32.8)	53.09 (54.04)
EDI_20050218-0900	13.0	7.1	1.9 (21.2)	21.99 (41.36)
EDI_20061113-1500	8.2	3.2	7.3 (5.5)	18.70 (16.89)
EDI_20061114-1500	2.6	4.4	0.8 (1.0)	7.78 (8.04)
NIST_20051024-0930	29.3	0.6	4.6 (3.9)	34.44 (33.77)
NIST_20051102-1323	22.7	3.5	3.3 (2.5)	29.53 (28.76)
NIST_20051104-1515	3.3	2.1	0.8 (0.4)	6.19 (5.85)
NIST_20060216-1347	2.0	2.5	4.5 (3.1)	9.04 (7.56)
VT_20050408-1500	0.8	5.0	2.9 (3.3)	8.76 (9.14)
VT_20050425-1000	4.9	3.0	1.4 (1.0)	9.34 (8.93)
VT_20050623-1400	15.5	10.7	15.9 (7.4)	42.05 (33.60)
VT_20051027-1400	11.7	7.5	16.8 (30.1)	36.03 (49.35)
ALL	12.6	4.1	6.3 (7.7)	23.01 (24.40)

Language Modelling



$$lkld_{AA} = Prob(x(0) \mid \Theta_A) \prod_{i=1}^{MD-1} (1 \cdot Prob(x(i) \mid \Theta_A)) \cdot \prod_{i=MD}^{2MD-1} (\alpha \cdot Prob(x(i) \mid \Theta_A))$$

$$lkld_{AB} = Prob(x(0) \mid \Theta_A) \prod_{i=1}^{MD-1} (1 \cdot Prob(x(i) \mid \Theta_A)) \cdot \frac{\beta}{M} Prob(x(MD) \mid \Theta_B) \prod_{i=MD+1}^{2MD-1} (1 \cdot Prob(x(i) \mid \Theta_B))$$

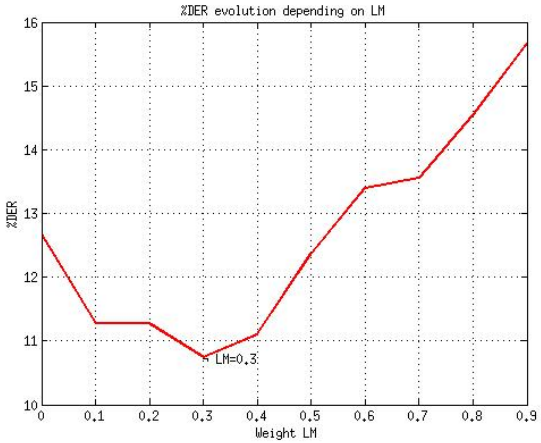
A speaker change occurs:

$$(1-LM)\cdot lkld_{AB} + LM \cdot \log(T_{AB}) > (1-LM)\cdot lkld_{AA} + LM \cdot \log(T_{AA})$$

The transition matrix T is normalized by the total number of transitions and LM is the weight

Language Modelling

- **Use of a language modelling**:
  - No trained LM. A transition matrix (T_a) is defined per each cluster. At each iteration T_a is updated with the information of the newly clustering
  - Count each transition among consecutive clusters



  - Count the trigrams of the form ABA and enforce the probability transition of speaker B. That's the common behavior in a conversation with short speaker interruptions
  - Depending on the length of this speaker interruptions the transition probability of the speaker B is increased (durations of 250ms and 150ms are taken into account)
  - Linear combination in log domain with the acoustics likelihoods
  - The LM is incorporated after the computation of the acoustics probabilities ensuring the Minimum Duration (MD) constrain

Evaluation results

MDM condition (SDM condition)

NAME SHOW	% MISS	% FA	% SER	% DER
EDI_20071128-1000	6.3(6.3)	2.8(2.8)	19.5(1.0)	28.53(10.03)
EDI_20071128-1500	8.5(7.0)	7.3(37.3	23.0(16.9)	38.83(61.17)
IDI_20090128-1600	5.2(3.7)	0.9(2.2)	10.6(7.4)	16.68(13.26)
IDI_20090129-1000	5.4(5.4)	6.8(6.8)	10.8(31.2)	23.01(43.44)
NIST_20080201-1405	15.0(15.0)	2.7(2.7)	49.4(47.3)	67.03(64.99)
NIST_20080227-1501	8.8(8.4)	1.0(2.8)	28.2(33.4)	38.09(44.58)
NIST_20080307-0955	4.5(4.5)	1.4(1.4)	26.9(22.0)	32.81(27.86)
ALL	7.1(6.5)	3.3(7.6)	21.6(20.3)	31.98(34.46)

System	MISS (%)	FA (%)	SER (%)	DER (%)
Contrast w/o LM-CV	7.1	3.3	43.4	53.73

Conclusions

The UPC RT09 Evaluation System

- RT 09 dataset is more difficult compared with previous datasets
- Shows with women's participation obtain a very poor performance. It suggests the using of gender dependent models or specific features
- The sdm'09 results are similar to those obtained in the mdm condition (some bug in the mdm condition? we need to check it)
- Future development: Working on realtime implementations. Tradeoff between algorithm performance/accuracy and realtime results

Thank you!!! Any question?